Event Study and Principal Component Analysis based on Sentiment Analysis – A combined methodology to study the stock market with an empirical study

Qianwen Xu[1], Victor Chang[1] and Ching-Hsien Hsu*[2, 3]

1. SCEDT Department of Computer Science & Information Systems, Teesside University, Campus Heart, Southfield Rd, Middlesbrough, TS1 3BX, UK
2. Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan
3. Department of Medical Research, China Medical University Hospital, Taichung, Taiwan

Emails: qianwen.ariel.xu@gmail.com; robertchh@gmail.com*

Abstract

This paper provides an improved method by introducing Sentiment Analysis into the Event Study and Principal Component Analysis. The model is constructed by using the heuristic mean-end analysis. This method enables us to take into investors' feelings towards related stocks when we study the stock market's reaction to a given event. This paper investigates the Chinese A-shared market over 2013 - 2019 to study the influence of rumors and the offsetting impact of rumor clarifications on the stock price. The results indicate that no matter investor sentiment is bullish or bearish, stock price reacts significantly to rumors before as well as when the rumor goes public. Furthermore, clarification offsets the positive abnormal returns caused by rumors with bullish sentiment substantially at a limited level. Still, after five days, it creates a positive effect like the positive rumor does on the stock price. Under the bearish sentiment, clarification brings an insignificant impact on the stock price. The results indicate that the source of rumor may not come from the media and investment decisions established

on rumors would be beneficial to investors before as well as after they are published. Moreover, official clarification causes an offset effect, but it is very limited.

Keywords: Sentiment analysis; Event Study; Principal Component Analysis; Rumor analysis; Stock.

1. Introduction

1.1 Background

For both regulating trading and the efficiency of stock markets, it is one of the significant issues that the stock markets can be manipulated (Aggarwal & Wu, 2003). Stock prices can be influenced in many ways, such as insiders manipulating enterprises' accounting and earnings, purchasing a large amount of stock, especially releasing rumors to the markets.

Rumors about stock markets can be seen nearly every day in the popular media. It is a common phenomenon in the Chinese stock markets as well and has been a chronic problem that hindered the development of stock markets. Therefore, this research attempts to study the Chinese A-shared market by concentrating on the effect of rumor release as well as the offsetting effect of rumor clarification.

Among the existing literature, Event Study is widely used to analyze a stock market's reaction to a given event. However, most studies fail to consider the impact of investors' sentiment on investment decisions, then on the target stock price. This paper provides an improved methodology combining Sentiment Analysis (SMA) and Event Study so that the investors' feelings can be considered.

In the area of building the stock price prediction model, many studies employ the multivariable regression but counter the multicollinearity problems. To deal with the problem, they have to remove significant variables at the expense of the explanatory strength of the model. In addition, they do not take the sentiment into account as well.

This paper selects the output of SMA as a variable of principal component regression (PCA), which is a method to solve the multicollinearity without reducing the number of variables. Hence, the multicollinearity problem can be solved, as well as the number of related variables can be maintained.

1.2 Research Questions and Reasons

This paper employs SMA, Event Study and PCA to study the following two research questions:

Q1: Will the event of rumors release cause abnormal return before or after publication?

Q2: Will the announcements on rumors offset the effect of rumor release on the stock?

Q1 focuses on the influence of releasing rumor through the media and Q2 focuses on the impact of clarifying rumor by related list firms. This research focuses on the two aspects for three reasons.

First, rumor release and clarification cause two symmetrical information flow and based on the efficient market theory (Malkiel and Fama, 1970) and noise deal(Black, 1986), the activity of clarifying rumors should offset the effect of releasing them on the markets. However, it seems that the theory is different from the facts. For example, on 25th February 2008, a rumor spread on Weibo that China Unicom would refinance and on that very day, China Unicom's stock price decreased by 11% (SSE,2008). On the following day, the company released the clarification and attempted to stop the decrease. However, the price kept sinking and dived by around 50 percent during four months. There are similar events all over the world. Therefore, it is meaningful to do this research in that it may challenge the everyday consumption that the activities of rumor clarification will offset the effect of rumor release.

Second, if the asymmetric impact is a common phenomenon, then it means that stock prices can be manipulated by rumors. If so, the manipulation may cause harm to investors' benefits, make it challenging to regulate market trading, reduce investors' confidence in the accuracy and transparency of market information (Zhao, He & Wu, 2010). Therefore, this research is important to investors and market regulators as well.

Third, research on the effect of rumors is limited in the Chinese market. According to the statistic from China National Knowledge Infrastructure (CNKI,2019), there have been only 14 papers related to the relationship between rumors and stock since the first one published from 2010 to the present. Among them, only six articles are also related to the offsetting effect of clarification.

1.3 Overview of Methods and Results

The model construction process is heuristic and uses mean-end analysis. After data collection, Sentiment Analysis will be employed to analyze and quantify investors' feelings about rumors. Then, the impact of rumor release and the announcement will be analyzed based on the Event Study method. Finally, a prediction model based on Principal Component Analysis will be built. Using the method of Sentiment Analysis allows us to identify a stock to be bullish or bearish by its investors' sentiments. This can be done without the authors' subjective attitude towards the related events and by introducing the sentiment as a factor into Event Study and PCA makes us be able to learn how the stock market reacts to a given event more deeply.

This paper finds that the rumors will affect related stocks before, as well as on the day it is released in the public media, but the offset effect of the clarification is minimal. The results provide the references for market regulators about unsymmetrical investor actions and the level of efficiency of the market.

Section 2 introduces existing literature on the research about rumors and stock price as well as on the methods this paper employs. Section 3 outlines the methods and models this paper conducts. Section 4 describes the process of data collection and management. Section 5 introduces the results and findings and Section 6 provides an evaluation of the results. Finally, Section 7 concludes this paper and underlines the contributions and limitations.

2. Literature Review

2.1 Rumors and stock price

The impact of rumors on stock markets is well investigated theoretically as well as empirically. Diefenbach (1972) and Logue and Tuttle (1973) conducted two initial studies and reported that investors could not take advantage of analysts' suggestions or rumors, no matter they are published or not. Diefenbach (1972) attempted to make an overall assessment of the value of recommendations received from brokerage firms and compared the performance of related stocks over 52 weeks to the Standard & Poor's index and found that the recommendations based on rumors has no value. Logue and Tuttle (1973) examined the performances of brokerage houses' investment suggestions and compared them with the randomly selected securities. He concluded that an investor who takes the advice of brokerage firms as a routine would do on balance as well by stochastically selecting securities.

However, later studies reported significantly different from the two initial studies. By investigating positive rumors from the 'Heard On The Street' column of ET on the Istanbul Stock Exchange, Kiymaz found that abnormal returns are positive and significant four days before the publication date, but abnormal returns are insignificant after the publication. The results among the two periods indicate that investors can benefit from the investment decisions founded on rumors, but the decisions would become worthless when they are published (Kiymaz, 2001). Different from Kiymaz, Lloyd-Davies and Canes (1978) studied the rumors posted on the Wall Street Journal's column, which is named 'Heard On The Street', and found that rumors do not only affect stock price significantly before it is published but also after. In 2015, Ahern and Sosyura found that media coverage of merger rumors shows a bias towards newsworthy firms that attractive to a broad audience. However, the reduced accuracy regarding newsworthiness is not reflected in the stock price (Ahern and Sosyura, 2015).

As for the offset effect of clarification, Yang and Luo investigated the influence of rumor clarification announcements on the stock returns under a bull market and a bear

market, respectively, from 2007 to 2011, in the Chinese market. They studied clarification about positive rumors and found that when the positive rumors are clarified, significantly positive average cumulative abnormal return is observed in a bull market. Still, the significantly negative abnormal return is detected in a bear market (Yang & Luo, 2014).

Unlike the influence of rumors, research about rumor clarification are very rare, especially in the Chinese stock market. Therefore, this paper also aims to study the offset effect of clarification.

2.2 Methods

Investors' sentiment can have an impact on their investment decisions. As is shown in Figure 1, when an event is reported by the media, the event will be interpreted by investors. Firstly, their translation will influence their sentiments and then they will decide to buy, sell, or hold the stocks. Secondly, the market will reflect the investors' behavior through its price movements (Li. et al., 2014). Therefore, it is meaningful and crucial to analyze the rumor's impact on the stock price from the sentiment aspect.

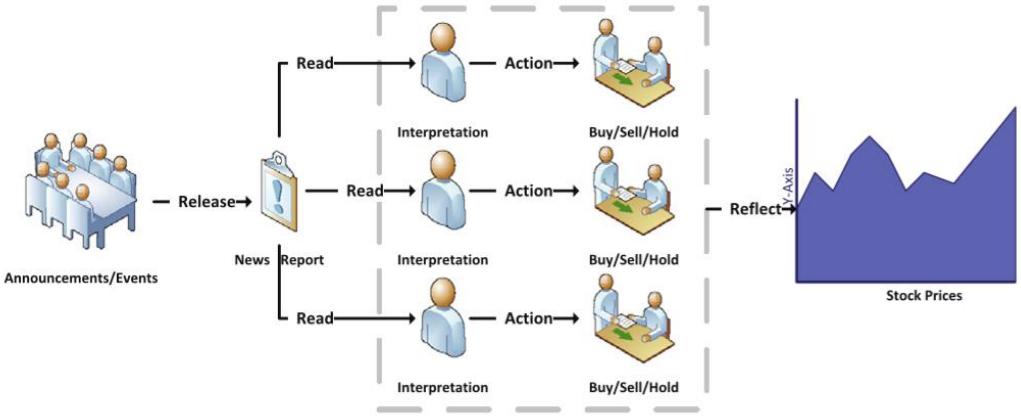

Figure 1. The general situation that events affect market prices ((Li et al., 2014)

2.2.1 Event Study

An Event Study is a method conducted on the stock or other securities that tests the impact of a significant event occurrence. It can disclose essential information on how protection is possible to respond to a given event (Fama, 1991). Among the existing literature, Event Study is widely used both in China and abroad in the area of studying the security. He (2015) adopts event studies to analyze the relationship between the effectiveness of clarification and the quality of clarification from three aspects of timeliness, detail and the wording of the clarification. Going one step further than He (2015), a recent paper focuses on the influence of the securities margin trading policy, which was introduced in 2007 in China, on the market's reaction to rumors and clarification announcements (Chen, 2017). However, both of them do not consider the impact of investors' sentiment on investment decisions, then on the target stock price. They assume that a stock's bullish mood or bearish mood is based on the content of the rumor instead of investors' feelings or moods towards the rumor. Their assumption is subjective as they classify the stocks by themselves when what the stock market reflects is all the investors' sentiments.

2.2.2 Stock Price Prediction Model

As for the methods used to build the stock price prediction model, Chen (2016) constructs her model based on the multiple variable regression. In her first model, she selects seven variables, but she finds a multicollinearity problem and she solve the problem by removing four variables from the model. Although she manages to deal with the multicollinearity problem by reducing the number of independent variables, the accuracy of her model also decreases, which is an important factor in prediction. To improve the accuracy of the multivariable regression model used in the stock price prediction, Teng and Zheng (2019) provide a multiple regression model based on deep learning (MRDL) in order to improve the accuracy. Similar to Chen's study, there are three variables in their model. However, there are two limitations in their studies. On the one hand, like He (2015) and Chen's (2017) studies, their studies do not take investors' sentiment as a variable into consideration, which is very significant in

prediction. On the other hand, a lot of variables can have impacts on stock price according to the existing literature. While aiming to overcome the multicollinearity problem, two studies have to reduce the explanatory variables in their models at the expense of the model's accuracy.

2.2.3 The Improved Methodology based on SMA

This paper provides an improved methodology combining Sentiment Analysis (SMA) and Event Study together and the output of SMA can be used as the input of Principal Component Analysis (PCA).

SMA is useful in classifying emotions. In the area of disaster relief, Beigi et al., (2016) described SMA in social networks and the advantage of applying it during the period of emergencies and disasters. Mittal and Goel (2012) employed SMA to investigate the causative relation between public mood collected from twitter.com and the DJIA values. They found that among the observed dimensions of feelings, only calmness and happiness are Granger causative of the DJIA by 3-4 days.

Instead of classifying the rumors based on the content by the researchers, SMA is used to analyze the investors' reviews, which express their feelings or emotions about the rumors and then classify the related stocks by these feelings. The stocks will be grouped by the feelings and then Event Study can be conducted on each group, respectively. The results will be more detailed and accurate. Additionally, this paper carries out the multivariable regression based on the principal component analysis. The output of SMA will be an independent variable in this regression as the sentiments will be quantified in SMA. This method can solve the multicollinearity problem as well as maintain the number of related variables.

3 Methodology

There are mainly three methods this paper will use in different parts. The model construction process is heuristic and uses mean-end analysis. In this paper, it means we divide our problems into several sub-goals, which are classifying sentiments, quantifying sentiments by SMA and introducing them into the Event Study and PCA to

conduct further numerical analysis. By achieving the sub-goals, the problems will be solved. As is shown in figure 2, rumor's reviews will be input to the sentiment analysis and the reviews will be classified to be positive or negative, and their related stocks will be classified as bullish or bearish accordingly. The classified stocks will be used in the second method, Event Study. During the sentiment analysis, the sentiment will also be quantified and the quantified emotions will be employed in the third method, PCA, to build a prediction model of the stock price as an independent variable.

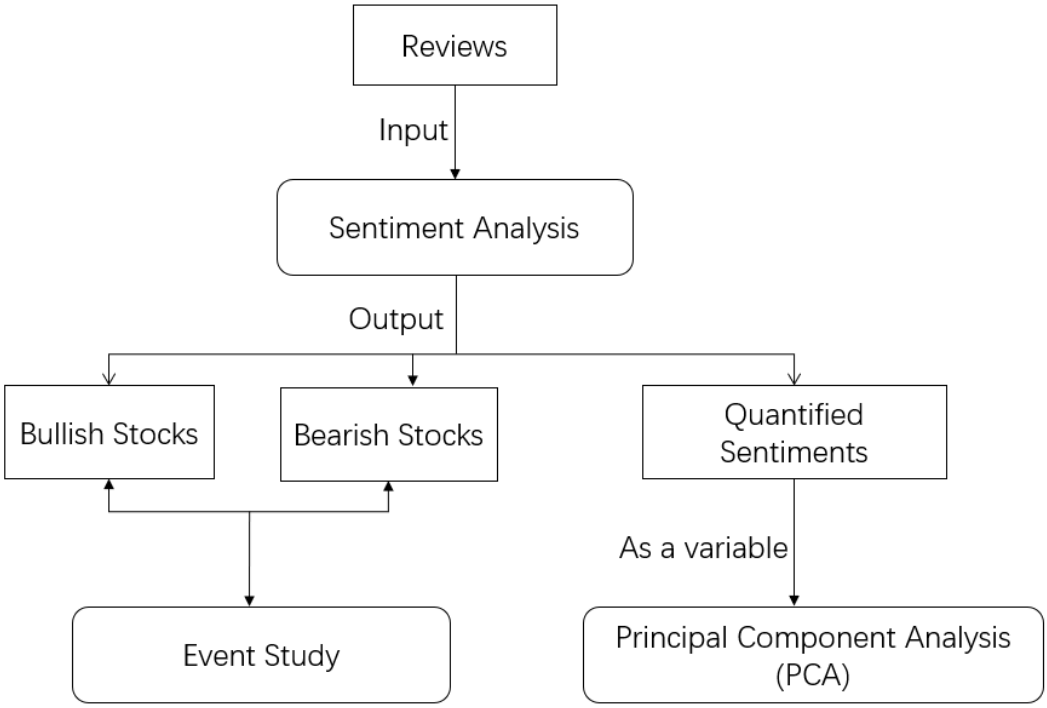

Figure 2. Methodology

3.1 Sentiment Analysis (SMA)

In sentiment analysis, the crucial task is to learn how texts or words can show a person's positive or negative attitudes towards a subject. (Nasukawa and Yi,2003). Therefore, sentiment analysis is to identify:

- Sentiment expressions;
- The polarity of the emotions;
- The relationship between them and the subject.

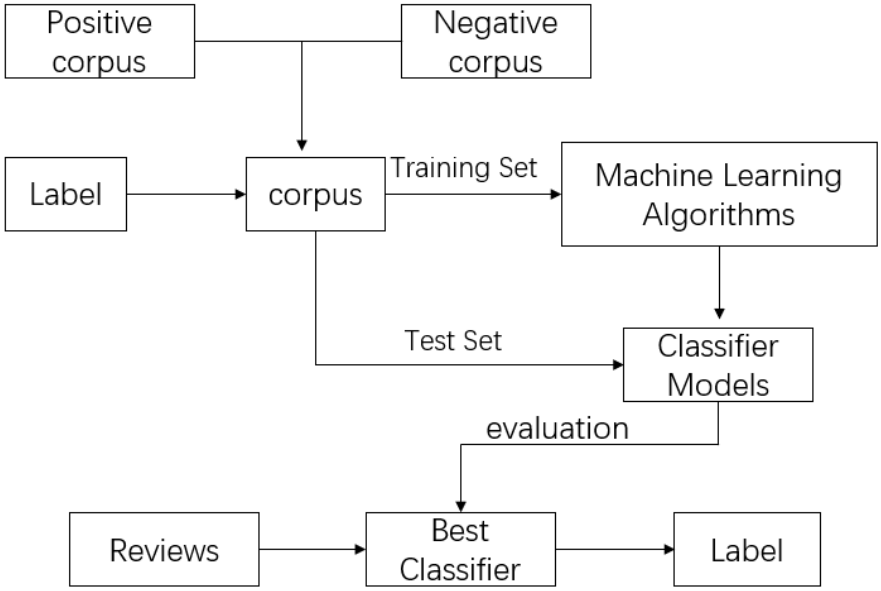

Figure 3. Process of Sentiment Analysis

In this paper, SMA will be conducted in Python. As is shown in figure 3, after preparing the positive and negative corpus, the corpus will be labeled manually and then divided into the training set and the test set. In order to guarantee the accuracy of the results, there are eight models of machine learning algorithms to be applied to the training set and then evaluated through the test set. The most accurate one will be used in the review classification. The algorithms are LinearSVC, LogisticReg, SGD, MultinomialNB, KNN, decision-trees, RandomForest and AdaBoost:

- Linear SVC (Support Vector Classifier)'s aim is to fit the data provided and return a "best fit" hyperplane that is able to divide or classify data. It has greater flexibility in choosing penalty and loss functions. This algorithm is suitable to be used on large numbers of samples (Fan et al., 2008).
- LogisticReg (Logistic Regression) is usually used to solve a 2-classes problem by modeling the probability of a sample belonging to one of two classes. It can also be used to classify multinomial types (Kleinbaum et al., 2002).
- SGD (stochastic gradient descent) is used to look for the minima of a function. SGDClassifier is a linear classifier which is suitable for data represented as dense or sparse arrays of floating-point values for the features (Bottou, 2010).

- MultinomialNB (Multinomial Naive Bayes) is one of Naive Bayes classification algorithms. Naive Bayes classifier is a common term that each feature in the model is conditionally independent. However, the Multinomial NB classifier employs a multinomial distribution for each feature (Su & Matwin, 2011).
- KNN (K-Nearest Neighbor) is used to categorize a data point. KNN is a non-parametric algorithm because no assumption is made on the data distribution. KNN can be used not only in the classification but also for regression (Peterson, 2009).
- Decision-tree builds classification or regression models by constructing tree structures. It divides the data set into smaller and smaller subsets during the process and an associated decision tree is gradually formed in the meantime. Finally, a tree with two different nodes is formed. One is decision nodes and the other one is leaf nodes. Each decision node connects to more than one branches and each leaf node stands for a classification. This algorithm is suitable to conduct sentiment analysis as it can process both categorical and numerical data (Quinlan, 1996)
- RandomForest is a learning algorithm similar to DecisionTree but is an ensemble method for classification. It constructs a vast number of decision trees and outputs the categories (Liaw & Wiener, 2002).
- AdaBoost is a learning method employed to improve the performance of any learning algorithm and in classification, DecisionTree is the one making use of AdaBoost most commonly. AdaBoost is suitable to classify the binary problems (Hastie et al., 2009).

Each review will be labeled to polarity as 1 or 0, which represents positive and negative, and then they will be grouped by their related stock number, the sentiment score of each stock can be calculated according to the equation (1):

bullishness index = $ln\left(\frac{1+M^{bull}}{1+M^{bear}}\right)$ (1)

where the bullishness index (BI) is the index to evaluate investors' sentiment when they are aware of related rumors. If the bullishness index is greater than 0, investors'

sentiment is bullish. Otherwise, investors' sentiment is bearish. Mbull is the number of stock reviews labeled as 1; Mbear is the number of stock reviews labeled as 0. (Oh and Sheng, 2011)

3.2 Event Study

With the aim to study how rumors and their clarification announcements affect the stock price through investors' sentiments, this paper will use the Event Study method.

An Event Study is widely used in the area of finance and economics. Kothari and Warner (2007) report that the number of event studies exceeds 500 and continues to grow. Therefore, it is suitable to employ the Event Study methodology in this paper.

Firstly, rumor release is defined to be Event 1 and the day it takes place is defined to be time t = 0; the activity of rumor clarification is defined to be Event 2 and the time t = 1 is the day it happens:

Event 1: rumor release;
Event 2: rumor clarification
Date of event 1: t=0;
Date of event 2: t=1.

Secondly, the event window will be defined to be ten days before event 1 and ten days after event 2. The estimation window will be defined to be 180 days before the event window:

Event window of event 1: t = [-10, 0];
Event window of event 2: t = [1, 11];
Estimation window: t = [-190, -11]

The event window is the period this paper focuses on and the data from the estimation window is used to predict the normal return rate. While using the trading data of 180 days in the estimation window, α_i and β_i of the following equation will be estimated for

each stock by using Ordinary Least Square Estimation (OLS) (Montgomery et al., 2012):

$ER_{i,t} = \alpha_i + \beta_i R_{m,t}$ (2)

$ER_{i,t} = ln\left(\frac{Price_{i,t}}{Price_{i,t-1}}\right)$ (3)

$R_{m,t} = ln\left(\frac{MPrice_t}{MPrice_{t-1}}\right)$ (4)

where $ER_{i,t}$ is the actual return rate for stock i for day t in the estimation window as well as the expected return rate for stock i for day t in the event window, $R_{m,t}$ is the expected rate of return for the market for day t; $Price_{i,t}$ is the close price of a stock for day t; $MPrice_{i,t}$ is the close price of the market for day t;

Then each stock's predicted normal return rate in the event window can be calculated by putting estimated α_i, β_i and $R_{m,t}$ in the event window into equation (2) and their abnormal return rate can be calculated according to the following equation:

$AR_{i,t} = R_{i,t} - ER_{i,t}$ (5)

where $AR_{i,t}$ is the abnormal return rate for stock i for day t and presents the difference between the actual return rate and predicted normal return rate. R_{it} is the actual return rate for stock i for day t; $ER_{i,t}$ is the predicted normal return rate for stock i for day t, which is calculated by equation (2).

Then, the averaged abnormal return and the cumulative abnormal return rate can be calculated according to the following equations:

$AAR_t = \frac{1}{n}\sum_{i=1}^{n} AR_{i,t}$ (6)

$CAR_t = \sum_{t=t_1}^{t=t_2} AAR_t$ (7)

where AAR_t is the averaged abnormal return and represents the sample average of each day's all stocks' abnormal returns in the event window. CAR_t is the sum of AAR_t from date t_1 to date t_2. Then n is the number of stocks in the sample (31 stocks in this paper).

Finally, to test whether the effects of rumor release and clarification are significant, the test statistics will be conducted. Test statistics are conducted to evaluate the degree of AAR's or CAR's significance in the event window. During the test, the t-values of AAR and CAR are calculated by equation (8)(9) and then compared with the critical value at a specific level of significance. If a t-value is greater than the critical value, the null hypothesis that 'the influence of the given event is not significant to the stock price' is rejected. (Asquith, 1983);

$$t_{AAR} = \frac{\mathbf{AAR}_t}{\frac{S(AAR_t)}{\sqrt{n}}} \quad (8)$$

$$t_{CAR} = \frac{\mathbf{CAR}_t}{\frac{S(CAR_t)}{\sqrt{n}}} \quad (9)$$

where S is the standard deviation of the abnormal return, in this paper, we run the t-test at the confidence level of 99%, the impact will be significant if the P-value is less than 0.01.

3.3 Principal Component Analysis (PCA)

Finally, this paper will use principal component analysis in SPSS to construct a prediction model for the stock price. PCA is based on the multivariable linear regression model and it is used because 16 variables will be chosen to analyze and too many variables may cause a multicollinearity problem, which means an exact or approximate linear relationship may exist between explanatory variables and this problem will lead to unreliable regression estimates (Mansfield and Helms, 1982). PCA enables to solve this problem by a reduction in the multidimensional data through fewer new variables (Lafi and Kaneene, 1992).

3.3.1 The process of PCA

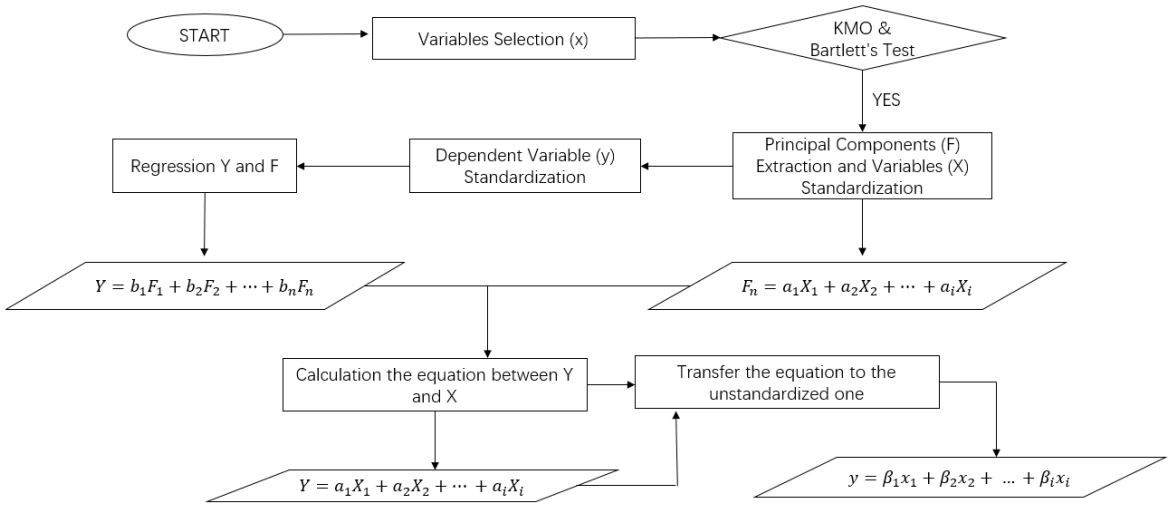

Figure 4. Process of Principal Component Regression

In the process of PCA, as is shown in Figure 4, the targeted data will be evaluated by the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMS) test. If the data passes the KMS test, several principal components (factors) will be extracted and standardized explanatory variables can represent these factors:

$$F_n = a_1X_1 + a_2X_2 + a_3X_3 + \cdots + a_iX_i \quad (10)$$

F_n: principal components;

X_i: standardized independent variable

After extracting out the five principal components, the relationship between the principal elements and dependent variables should be regressed again. However, during the extraction, the explanatory variables have been standardized while the dependent variable has not. Therefore, the dependent variable needs to be normalized first. After the second multivariable linear regression, the regression equation between standardized dependent variable and factors will be obtained:

$$Y = b_1F_1 + b_2F_2 + b_3F_3 + \cdots + b_nF_n \quad (11)$$

Finally, the standardized equation between variables can be obtained by the two equations and the unstandardized equation can be transformed from the standardized equation:

$$Y = c_1X_1 + c_2X_2 + c_3X_3 + \cdots + c_iX_i \quad (12)$$

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ \ldots + \beta_ix_i \quad (13)$$

x_i: unstandardized independent variable;

y: unstandardized dependent variable;

3.3.2 PCA outputs

PCA is based on multivariable regression and the linear regression will be conducted on the variables before the factor analysis. Outputs will include the following:

- R^2 and adjusted R^2 are to test the goodness-of-fit for a linear regression between a dependent variable and independent variables. This value shows the percentage of the variance in the dependent variable that can be explained by the independent variables jointly. Adjusted R^2 is a better indicator as it applies a penalty for increasing parameters while R^2 does not.
- The standard error of the estimate is used to determine how well a regression line can fit a data set. The value will be higher if the influence of random changes is significant.
- Durbin-Watson (DW) indicates an autocorrelation problem, which arises if different error terms are correlated. The value's range is from 0 to 4. 0 indicates positive autocorrelation and 4 indicates a negative correlation. If DW is close to 2, there is no autocorrelation problem.
- P-value is computed on the basis of the assumption that the difference in the sample is random. It is used to test whether the relationship between one explanatory variable and the dependent variable is significant. If the significant level is 95%, the null hypothesis is rejected if the p-value is less than 0.05.

- Unstandardized beta (B) is the slope of the line between the explanatory variables and the dependent variable. One unit increase in one independent variable will cause B units to increase in the dependent variable.
- The standard error for the unstandardized beta (SE B) shares a similar meaning with the standard error of the estimate. The higher the value, the less likely the relationship is significant.
- Standardized beta (Beta) is similar to a correlation coefficient. The value varies from -1 to 1. The relationship is more significant if the absolute value is closer to 1.
- T-test statistic (t), is calculated for the individual predictor variable. The higher the t-value is, the more significant the relationship is.
- The Variance Inflation Factor (VIF) is used to detect multicollinearity, which arises when an exact or approximate linear relationship exists between independent variables, resulting in unreliable regression estimates. Generally, if the value of VIF is 10 or more, the multicollinearity problem may exist.

By regressing the explanatory variables with the dependent variable, this paper can identify the multicollinearity problem and then carry out the factor analysis. Main outputs of factor analysis include the following:

- Kaiser-Mayer-Olkin Measure of Sampling Adequacy (KMS) is to measure whether the dataset is suitable to conduct the factor analysis and the critical value is 0.5. If KMS is greater than 0.5, the dataset is appropriate to do the analysis.
- p-value (Sig.): a measure indicates whether the data passes the KMS test. If the p-value is less than 0.05, the data passes the test.
- Initial Eigenvalues: The principle of principal component number extraction is the first m principal components whose eigenvalues corresponding to the principal component are greater than 1. The eigenvalue can be regarded as an indicator of the strength of the principal component.
- % of Variance explains the percentage of information the factor contains.

- Cumulative % explains the increasing percentage of information contained by the extracted factors.

4. Data

4.1 Data Collection

There are four datasets this paper will collect, including daily trading data, financial data, clarification announcements and reviews related to the rumors and clarification announcements from 2013 to 2019.

The first three datasets are from the China Stock Market & Accounting Research Database (CSMAR) and the fourth dataset is from Eastmoney.com. Daily trading data is used in the Event Study and financial data is employed in the part of building a prediction model based on PCA. Clarification announcements data is used to collect the reviews of rumors.

The reason why this paper looks into rumors by referencing clarification announcements is as follows. In 2007, the China Securities Regulatory Commission (CSRC) issued a regulation that all public companies on the Shanghai Stock Exchange (SHSE) and the Shenzhen Stock Exchange (SZSE) are required to deliver a clarification timely when they are involved in rumors. This regulation offers a great approach to study the effect of rumor and its clarification on the stock price.

The reviews of the rumor are collected from Eastmoney.com, one of the most influential websites among individual investors, and nearly every related news will be reprinted to the website in time.

To obtain reviews of rumors, the mean-end analysis is used as well and the steps are as follows:

First, after obtaining the clarification announcements dataset, this paper will filter out the stocks that their suspend reasons have the keyword 'clarification.' The reasons with the word 'clarification' mean that there are important events that occurred about the list companies and they have to clarify them.

Next, the clarification announcements' content will be obtained from the website of SHSE and SZSE because the dataset only has reasons but lacks content. Then, the rumors can be found accordingly.

Finally, according to the clarification announcements and rumors on Eastmoney.com, the rumors' spread date, investors' reviews about the rumors will be collected by using a data-crawling software named ‘bazhuayu’ online.

4.2 Data Cleanness

There are several selection rules required to be satisfied for a stock with a rumor to be included in the final sample. The rules are listed below:

1. There is a specific date in the clarification announcement. If a firm does not include a date of rumor release or the date is vague, the announcement should be excluded from the sample.
2. The rumor is confirmed to be fake or inaccurate in the official clarification announcement.
3. Each announcement is related to only one rumor. If an announcement is related to several rumors and the firm clarifies some of the rumors while others are proven to be accurate, the relevant stock will be removed from the sample.
4. The firm must have at least 190 trading days before the date of rumor release and ten days after the date of the announcement to meet the requirements of Event Study in section 3.2.
5. The rumor has reviews on Eastmoney.com. If there is no comment about the related rumor, the rumor should be excluded from the sample.

If a sample does not meet the criteria, then it should be excluded. After the selection progress, there are 31 stocks left.

5. Results

5.1 Sentiment Analysis Results and Analysis

Before training the sentiment classifiers, the reviews are tokenized first as the Chinese sentences do not have spaces between each word. The result of the tokenization is in the appendices. During the process of building a sentiment classifier, eight algorithms are evaluated and the result is as follows:

Table 1 Accuracy Result of Classification Algorithms

Algorithm	LinearSVC	LogisticReg	SGD	MultinomialNB
accuracy	0.8825	0.8809	0.8816	0.8796
Algorithm	KNN	DecisionTree	RandomForest	AdaBoost
accuracy	0.8209	0.7966	0.8271	0.7719

As is shown in the above table, Linear SVC (0.8825), LogisticReg (0.8809), SGD (0.8816) and MultinomialNB (0.8796) perform substantially and the value of their accuracy is close to each other, which is 0.88. But LinvearSVC is selected according to its highest value to classify each rumor review's sentiment in the next step of classification.

Then, the classified reviews will be grouped by their assigned numbers and then the bullishness index of each stock will be calculated by the equation (1) described in section 3.1. The result is as follows:

Table 2 Bullissness Index of Stocks – Positive

stock number	Bullishness Index
1_r	0.095310
2_r	0.356675
4_r	0.356675
5_r	0.182322
9_r	0.470004
10_r	0.087011
11_r	0.405465
15_r	0.211309
17_r	0.693147
20_r	2.397895
21_r	0.167054
22_r	0.095310
24_r	0.693147
25_r	0.693147
26_r	0.693147
27_r	0.122602
30_r	0.182322

Table 3 Bullissness Index of Stocks – Negative

stock number	Bullishness Index
3_r	-0.38777
6_r	-1.79176
7_r	-0.28768
8_r	-1.01160
12_r	-0.18232
13_r	-1.18626
14_r	-0.33647
16_r	-0.19106
18_r	-1.79176
19_r	-0.18232
23_r	-0.04725
28_r	-0.18232
29_r	-0.15415
31_r	-1.01160

This paper divides the results into two groups, as in table 2 and table 3. Among the 31 stocks, there are 17 stocks' BI value is greater than 0, meaning that the investors' sentiment is positive and consider the stock price will increase, and there are 14 stocks' BI value is less than 0, which indicates that the stocks' prices are deemed to go down.

5.2 Event Study Results and Analysis

After identifying investors' sentiments about the rumors, this paper is able to study how rumors or its clarification can influence stock price through investors' sentiment towards these stocks. Abnormal return rate calculation is in section appendices. The

result is divided into two groups. One is the stocks with bullish sentiment and the other one is the stocks with the bearish sentiment. The result summary and trend are as follows:

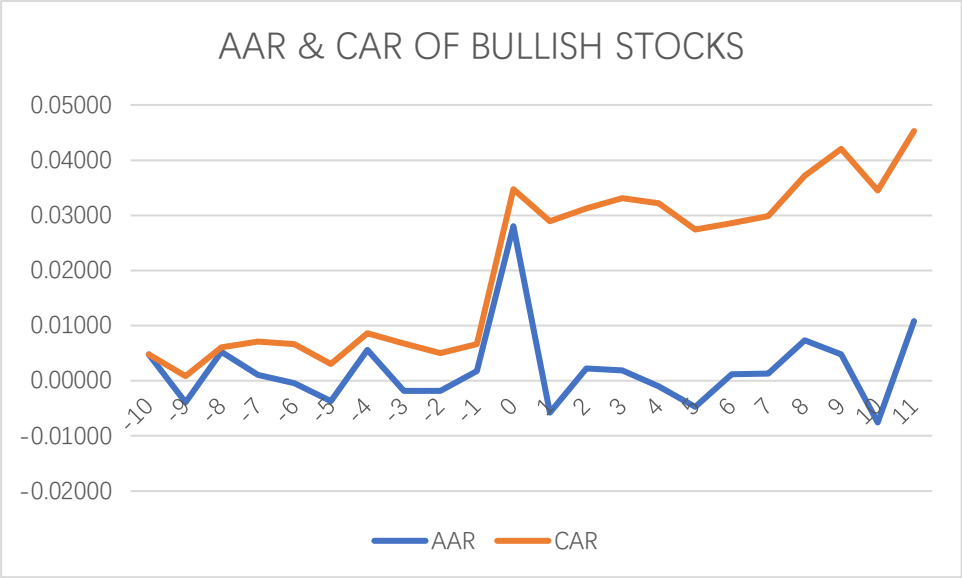

Figure 5. AAR & CAR OF BULLISH STOCKS

Table 4 T-test of Bullish Stocks

t	AAR	P-V	CAR	P-V	t	AAR	P-V	CAR	P-V
-10	0.00479	-	0.00479	-	1	-0.00573	-	0.02897	-
-9	-0.00396	0.279	0.00083	0.126	2	0.00225	0.254	0.03122	0.072
-8	0.00521	0.030	0.00604	0.016	3	0.00192	0.026	0.03314	0.012
-7	0.00108	0.012	0.00712	0.008	4	-0.001	0.011	0.03214	0.005
-6	-0.00046	0.008	0.00666	0.005	5	-0.00475	0.008	0.02739	0.005
-5	-0.00364	0.007	0.00302	0.004	6	0.00118	0.006	0.02857	0.004
-4	0.00559	0.006	0.00861	0.004	7	0.00127	0.005	0.02984	0.003
-3	-0.00186	0.005	0.00675	0.003	8	0.00737	0.005	0.03721	0.004
-2	-0.00179	0.004	0.00496	0.003	9	0.0048	0.005	0.04201	0.005
-1	0.00173	0.004	0.00669	0.002	10	-0.00756	0.005	0.03445	0.005
0	0.02802	0.009	0.0347	0.009	11	0.01082	0.005	0.04527	0.005

Interpretations of the results of stock prices with the bullish sentiment:

Event window of event 1:

- AAR and its p-value: At day 0, when the rumor is released, the average abnormal return has a sharp rise. Additionally, the p-value has been less than 0.01 since day -6, indicating that AAR passes the t-test six days before the day the rumor goes public.
- CAR and its p-value: The trend of CAR is positive from event day -10 to day 0. At the event day 0, the value reaches to its highest. Additionally, the p-value starts to be less than 0.01 from event day -7, meaning that CAR passes the t-test seven days before the day of rumor released.

Event window of event 2:

- AAR and its p-value: At day 1, when the rumor clarification is released, AAR decreases from 0.028 to -0.0075. The change of the value shows that clarification harms the stock price, trying to offset the effect of the rumor. But the value struggles during the event window and start to be positive again from day 6. The p-value shows the significant since day 5, which is less than 0.01.
- CAR and its p-value: Since the rumor clarification has been released, CAR decreases from day 1 to day 5, but it goes up again from day 6. And the p-value shows that on day 4, CAR starts to pass the t-test, meaning that the clarification is significant to the stock price from day 4.

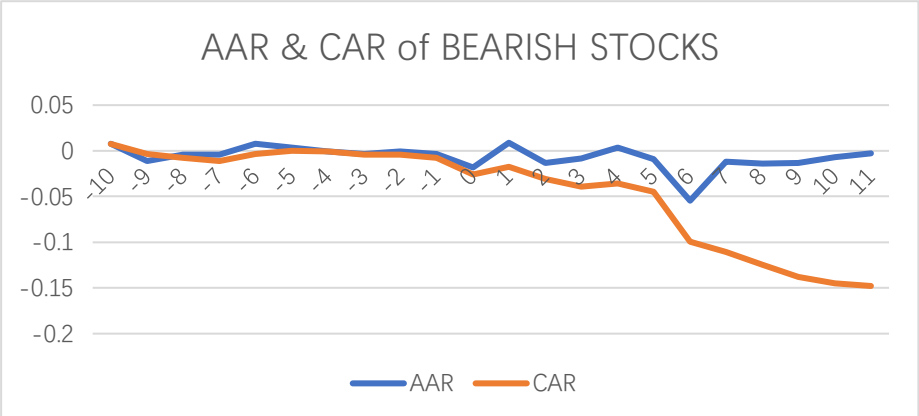

Figure 6. AAR & CAR OF BEARISH STOCKS

Table 5 T-test of Bearish Stocks

t	AAR	P-V	CAR	P-V	t	AAR	P-V	CAR	P-V
-10	0.00758	-	0.00758	-	1	0.00886	-	-0.01725	-
-9	-0.01103	0.592	-0.00345	0.177	2	-0.01327	0.704	-0.03052	0.195
-8	-0.00408	0.054	-0.00753	0.026	3	-0.0086	0.067	-0.03912	0.036
-7	-0.00385	0.023	-0.01139	0.013	4	0.00343	0.030	-0.03569	0.016
-6	0.00802	0.017	-0.00337	0.008	5	-0.00892	0.019	-0.04461	0.014
-5	0.00372	0.013	0.00036	0.010	6	-0.05452	0.037	-0.09913	0.013
-4	-0.00085	0.010	-0.00049	0.011	7	-0.01168	0.029	-0.11081	0.012
-3	-0.00341	0.008	-0.0039	0.010	8	-0.0138	0.023	-0.12461	0.012
-2	-0.00052	0.007	-0.00442	0.009	9	-0.01341	0.020	-0.13802	0.013
-1	-0.00329	0.006	-0.00771	0.009	10	-0.00722	0.017	-0.14524	0.013
0	-0.0184	0.007	-0.02611	0.008	11	-0.00264	0.015	-0.14788	0.014

Interpretations of the results of stock prices with the bearish sentiment:

Event window of event 1:

- AAR and its p-value: The trend of AAR is negative and reaches its lowest (-0.0184) when the rumor is released on day 0. CAR passes the t-test on day -4 with a p-value of 0.01, meaning that the negative abnormal return is related to the rumor four days before the rumor goes public.
- CAR and its p-value: At the event day 0, when the rumor goes public on the media, CAR decreases from -0.008 to -0.026, which is 2.25 times lower. The p-value starts to keep passing the t-test from day -3, meaning that stock price reacts statistically significant to the rumor.

Event window of event 2:

- AAR and its p-value: AAR becomes positive from -0.0184 to 0.0087 on the day the clarification announcement comes out, which is day 1. But From day 2, AAR turns negative again and remains negative. The p-values from day -10 to day 0 are all greater than 0.01, meaning that the effect of rumor clarification is not significant to the stock price at all.
- CAR and its p-value: AAR increase by around 40% (from -0.02611 to -0.01725) when the clarification is announced. But similar to AAR, it decreases again from day two and keeps going down. Additionally, the p-value indicates that CAR fails to pass the t-test all the ten days after the clarification, meaning that the offset effect is not significant to the stock price at all.

5.3 Principal Component Analysis

Finally, this paper will use principal component analysis in SPSS to construct a prediction model for the stock price.

5.3.1 Variables Selection

The evaluation should include the factors that are currently able to explain the change of stock price so that the results of this analysis can provide practical, feasible and useful recommendations for improving the supervision of Chinese stock markets. Based on the reference to other literature and in consideration of the availability of data, this paper chooses 16 factors as independent variables to study their relationship with the stock price. Aside from the sentiment index and stock market return, the other 14 variables are divided into four categories. They are solvency, profitability, activity and situation of capital structure. The summary of independent variables is described as below and their detailed explanation is in section appendices:

Table 6 Variables Description

Independent Variable	FullName	Variable Name
x_1	Sentiment	Sentiment
x_2	Current Ratio	CR(%)
x_3	interest coverage ratio	ICR(%)
x_4	income growth rate	IGR(%)
x_5	Return on equity	ROE(%)
x_6	earnings per share	EPS(RMB)
x_7	Cash flow ratio	CFR(%)
x_8	Return on Assets	ROA(%)
x_9	Equity multiplier	EM
x_{10}	Total asset turnover rate	TURNTA(times)
x_{11}	Quick ratio	QR(%)
x_{12}	Book Value of Equity per share	BVEPS (RMB)
x_{13}	Net Cash Flows Per-share generated by operating activities	NCFPSGO (RMB)
x_{14}	Asset-liability ratio	D/A(%)
x_{15}	Account Receivable Turnover Rate	TURNAR(times)
x_{16}	Market Rate	Rm(%)

5.3.2 PCA Results and Analysis

Because of the missing data of one stock, this paper excluded it from the samples in this part and there are 30 samples left. This paper divided the samples into two sets. One is the training set with 25 samples and the other one is the test set with five samples. The results are as follows:

Table 7 Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson	F	Sig.
1	.951[a]	0.905	0.715	12.04148	2.408	4.766	.016[b]

Interpretation of the key statistics in table 7:

- R^2 is 0.905 and adjusted R^2 is 0.715, indicating that the fitness of the model is not bad. Adjusted R^2 indicates that 16 variables together can explain the variance in the dependent variable's 'close price' by 71.5%.
- The standard error of the estimate is 12.04148 and is a little high in considering the data, which suggests that the effect of random changes is significant.
- The value of Durbin-Watson is 2.408 and is close to 2, indicating that the autocorrelation in the residuals of the regression is not the problem.
- P-value is 0.016, which is less than 0.05, meaning that the null hypothesis 'the coefficients of all the 16 variables are equal to 0' is rejected and all the explanatory variables together have a strong relationship with the dependent variable 'close price.'

Table 8 Coefficients

Model		Unstd. B	Std. Beta	Sig.	VIF
1	(Constant)	8.886		0.787	
	Sentiment x_1	-11.652	-0.262	0.238	3.544
	CR(%) x_2	0.368	0.046	0.939	29.030
	ICR(%) x_3	-0.008	-0.501	0.112	6.628
	IGR(%) x_4	0.113	0.286	0.117	2.238
	ROE(%) x_5	-28.483	-0.500	0.327	19.369
	EPS(RMB) x_6	0.764	0.601	0.109	9.345
	CFR(%) x_7	0.189	0.043	0.892	7.934
	ROA(%) x_8	6.212	0.312	0.581	24.743
	EM x_9	-2.963	-0.139	0.801	23.997
	TURNTA(times) x_{10}	0.000	-0.241	0.900	289.460
	QR(%) x_{11}	-1.411	-0.160	0.732	17.255
	BVEPS (RMB) x_{12}	-1.619	-0.148	0.776	21.368
	NCFPSGO (RMB) x_{13}	0.057	0.028	0.929	8.071
	D/A(%) x_{14}	0.249	0.232	0.748	41.234
	TURNAR(times) x_{15}	0.000	0.546	0.777	292.647
	Rm(%) x_{16}	315.285	0.261	0.102	1.689

When the 16 explanatory variables are analyzed respectively, the results are shown in table 8 and the interpretation is as follows:

- For x_4, its B is 0.113. It means for every 1% increase in income growth rate, the stock price increase by 0.113 RMB. For x_{11}, its B is -1.411. It means that the stock price decreases by 1.411 RMB as every 1% increase in a quick rate.
- Among the variables, x_6's Beta is the largest (0.601), meaning that it has the most significant relationship with 'close price' while x_{13}' relationship is weakest (0.028).
- All the variables' p-values are greater than 0.05. It seems like their relationships, respectively, with the dependent variable 'close price' are not significant.
- Among the 16 variables, 9 of their VIF are greater than 10. It indicates that an exact or approximate linear relationship may exist between independent variables and the coefficient of the variables are biased.

In order to solve the multicollinearity problem, this paper uses SPSS 25 to analyze the principal components and the outcome is as follows:

Table 9 KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	0.503
Bartlett's Test of Sphericity (Sig.)	0.000

The explanations are as follows:

- KMS is 0.503, which is higher than 0.5, meaning that the dataset this paper chose passes the KMO and Bartlett's Test and is suitable to conduct the factor analysis.
- Possibility of significance is around 0, which is less than 0.05 at the significant level of 95%, indicating that the explanatory variables are related to each other and confirms that the data is suitable to conduct the principal component analysis.

Table 10 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.483	28.018	28.018	4.483	28.018	28.018
2	3.856	24.099	52.118	3.856	24.099	52.118
3	1.949	12.18	64.298	1.949	12.18	64.298
4	1.596	9.973	74.271	1.596	9.973	74.271
5	1.027	6.419	80.69	1.027	6.419	80.69
6	0.885	5.534	86.224			
7	0.792	4.948	91.172			
8	0.44	2.751	93.923			
9	0.436	2.727	96.65			
10	0.197	1.228	97.878			
11	0.192	1.2	99.078			
12	0.085	0.528	99.606			
13	0.034	0.215	99.821			
14	0.016	0.099	99.921			
15	0.011	0.068	99.989			
16	0.002	0.011	100			

As is shown in table 10, the method of Principal Component Analysis extracts five principal components (F1, F2, F3, F4 and F5) and the percentage of their contribution to explain the dependent variable is about 80.69% cumulatively. The further interpretation is as follows:

- Initial Eigenvalues: Among 16 initial eigenvalues, the value of the first five components is greater than 1, meaning that their explanatory strength is better than the original variables. Therefore, they are extracted to be the principal components.

- The value of '% of Variance' shows that F1 explains 28.018% of the information the original explanatory variables contain and F2 explains 24.099%, F3 explains 12.18%, F4 explains 9.973% and F5 explains 6.419%.
- The value of 'Cumulative %' explains that the five principal components together explain 80.69% of the information the data contains.

Table 11 Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Sig.
1	0.951	0.905	0.715	12.04148	0.016
2	0.873	0.762	0.7	0.547943	0.000

After conducting the second linear regression (model 2), this paper compares it with model 1. The result is shown in table 11 and explanations are as follows:

- Model 2's R^2 is 0.762, 0.143 smaller than Model 1's and adjusted R^2 is 0.7, 0.015 lower than Model 1's. It means that model 2's goodness of fit is weaker than model 1's.
- Model 2's standard error of the estimate is 0.547943 and is much lower than model 1's, which is 12.04148. It means that model 2 has lesser points spread out from the regression line.
- Model 2's p-value is close to 0. It means that the null hypothesis 'the coefficients of all the five factors are equal to 0' is rejected and the relationship between the factors and dependent variables is significant. Furthermore, model 2's p-value is much smaller than model 1's, meaning that model 2 constructs a stronger relationship than model 1.

Although R^2 and adjusted R^2 of model 2 is lower than model 1's, the value of the standard error of the estimate and p-value shows that model 2 is better than model 1.

Table 12 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics
		B	Std. Error	Beta			VIF
2	(Constant)	2.05E-16	0.11		0	1	
	F1	0.047	0.053	0.099	0.881	0.389	1
	F2	0.078	0.057	0.153	1.364	0.189	1
	F3	0.523	0.08	0.73	6.527	0	1
	F4	-0.344	0.089	-0.435	-3.891	0.001	1
	F5	-0.083	0.11	-0.084	-0.751	0.462	1

The detailed results of the relationship between the five factors and the dependent variable are shown in table 12. Interpretation of the statistics is shown as follows:

- F1’s B is 0.047 and it means that for every 1 unit increase in F1, the stock price increase by 0.047 RMB. For F4, its B is -0.344. It means that the stock price decreases by 0.344 RMB as every 1 unit increase in F4.
- Among the factors, the absolute value of F4's Beta is the largest, which is 0.435, meaning that it has the most reliable relationship with the dependent variable 'close price' while F5's relationship is weakest with the smallest value of 0.084.
- F3, F4’s possibilities of significance are less than 0.05 while others’ are not, meaning that the relationship between the dependent variable ‘close price’ and F3, F4 is significant while the hypothesis that the coefficient of other factors equals 0 should not be rejected. Therefore, among the five factors, only F3 and F4 should be used to describe their relationship with the dependent variable ‘close price'.
- All the factors' VIF values are equal to 1, indicating that the multicollinear relationships between independent variables are not significant anymore. This confirms that model 2 is better than model 1.

6. Evaluation

6.1 Sentiment Analysis and Event Study

By introducing SMA into the Event Study, this paper can analyze the effect of rumor release or clarification announcements on stock prices through investors' sentiments.

On the one hand, by comparing bullish stocks and bearish stocks, this paper can find that they behave differently both in the event window of rumor release and clarification announcement.

During the event window of rumor release, bullish stocks' CAR passes the t-test seven days before the day the rumor goes public while bearish stocks' CAR surpasses the t-test just three days before the day. It seems that rumors with bullish sentiment influence the stock price much earlier than those with bearish emotions.

During the event of clarification announcement, the bullish stock price goes down after the announcement, but the offset effect is limited. The t-test shows that the clarification is not significant to the price until three days after the clarification. Moreover, although CAR passes all the t-test from day 4 to day 10, it can be seen that the event period can be further divided into two stages according to the change of CAR. One stage is from day 4 to day 5, since the CAR goes down, the significant effect belongs to the offsetting effect. The other one stage is from day 6 to day 11, the CAR stops going down but rebounding, meaning that the clarification causes a positive effect on the stock price rather than the negative one which it intends to do. As for the bearish stocks, the price rebounds a little when the clarification is announced. However, it stops increasing on the second day but decreases again. Additionally, both AAR and CAR fails all the t-test during the event of clarification announcement, meaning that the offset effect is not significant to the stock price at all. From the discussion above, this paper finds that despite the offset effect, the clarification caused is limited for both bullish and bearish stocks, it has greater impacts on bullish stocks than on bearish stocks.

On the other hand, from the perspective of the two events of rumor release and clarification announcement, no matter the investors' sentiment towards the related

rumors is bullish or bearish, there are consequences as follows. First, the rumor has already caused an effect on the stock price before the day the rumor goes public as the AAR and CAR pass the t-test, meaning that the event is significant to the stock price. This result indicates that the source of rumor may not come from the media. Second, although the market has already reacted to the rumor before it goes public, the CAR still changes in the same direction as it does on the day before the rumored release and passes the t-test. This outcome is the same as Lloyd-Davies and Canes 's study on the American stock market but not consistent with the findings of Kiymaz's research on Istanbul's stock market.

Third, the event of clarification has an offsetting effect on the stock price. Still, the result is minimal and the stock price will keep increasing if the sentiment is bullish or keep going down if the sentiment is bearish right after the offset effect disappears. It seems that investors have doubts about the facticity of the public companies' clarification and intend to ignore them. There is an old saying in China, 'better to think the worst and be pleasantly surprised.' The investors tend to believe the rumors and are skeptical about the clarification announcements. Therefore, the announcements cannot remove their expectations to bullish stocks and dispel their fear of bearish stocks, resulting in the limited offset effect and stock prices' ignorance of the clarification. The reason why investors choose this kind of strategy is likely because of two main reasons. Firstly, investors are not rational. Their irrational enthusiastic public attention (Huberman and Regev, 2001) makes them unable to identify the truth through rumors. Secondly, it is the trust issues between investors and public companies on information disclosure.

6.2 Principal Component Analysis

According to the statistics, the equation used to describe the relationship between the principal components and the explanatory variables can be figured out:

$$F3 = -0.08983X_1 + 0.0960X_2 - 0.1060X_3 + 0.5521X_4 - 0.1857X_5 + 0.0005X_6 - 0.2691X_7 - 0.1111X_8 - 0.1408X_9 + 0.3398X_{10} + 0.1196X_{11} - 0.2584X_{12} - 0.1895X_{13} - 0.0427X_{14} + 0.3383X_{15} + 0.4229X_{16} \quad (14)$$

$$F4 = 0.3613X_1 + 0.1349X_2 + 0.0226X_3 - 0.1561X_4 + 0.4662X_5 - 0.1693X_6 - 0.1891X_7 + 0.4899X_8 - 0.3696X_9 + 0.1268X_{10} + 0.1476X_{11} - 0.0741X_{12} - 0.1613X_{13} - 0.2891X_{14} + 0.1166X_{15} - 0.0378X_{16} \quad (15)$$

where

the coefficient=the coefficient of component/ (Initial Eigenvalues)1/2

Then, after regressing the standardized dependent variable and significant principal components, the equation on the relationship between them is:

$$Y = 0.5229F3 - 0.3444F4 \quad (16)$$

Then this paper plugs equation (14)(15) into equation (16), then equation (17) which represents the relations between the standardized dependent variable and standardized explanatory variables is as follows:

$$Y = -0.1714X_1 + 0.0037X_2 - 0.0632X_3 + 0.3425X_4 - 0.2577X_5 + 0.0586X_6 - 0.0756X_7 - 0.2269X_8 + 0.0537X_9 + 0.1340X_{10} + 0.0117X_{11} - 0.1096X_{12} - 0.0435X_{13} + 0.0773X_{14} + 0.1367X_{15} + 0.2341X_{16} \quad (17)$$

The final unnormalized equation can be obtained according to equation (18)(19):

$$\mathrm{Y} = \frac{\mathrm{y}-\bar{\mathrm{y}}}{\mathrm{S_y}} \quad (18)$$

$$X_i = \frac{x_i - \bar{x_i}}{s_{x_i}} \quad (19)$$

the value of $\bar{\mathrm{y}}$, $\mathrm{S_y}$, $\mathrm{s_{x_i}}$, xi is from Table 13, and the final equation is as below:

$$y = 28.8203 - 7.6347x_1 + 0.0296x_2 - 0.0009x_3 + 0.1352x_4 - 14.6691x_5 + 0.0745x_6 - 0.3306x_7 - 4.5210x_8 + 1.1430x_9 + 0.0001x_{10} + 0.1028x_{11} - 1.1969x_{12} - 0.0873x_{13} + 0.0827x_{14} + 0.00004x_{15} + 282.4532x_{16} \quad (20)$$

Table 13 Descriptive Statistics

Variables	Mean	Std. Deviation	N
ClosePrice(RMB) y	18.0796	22.5618	25
Sentiment x_1	0.5600	0.5070	25
CR(%) x_2	2.5468	2.8513	25
ICR(%) x_3	935.7428	1504.9496	25
IGR(%) x_4	20.0608	57.1682	25
ROE(%) x_5	0.5224	0.3963	25
EPS(RMB) x_6	26.6640	17.7333	25
CFR(%) x_7	4.2528	5.1582	25
ROA(%) x_8	1.6700	1.1322	25
EM x_9	2.1682	1.0601	25
TURNTA(times) x_{10}	29209.2800	55663.4790	25
QR(%) x_{11}	1.9632	2.5659	25
BVEPS (RMB) x_{12}	1.8676	2.0660	25
NCFPSGO (RMB) x_{13}	10.0296	11.2431	25
D/A(%) x_{14}	45.0668	21.0937	25
TURNAR(times) x_{15}	36223.8800	69574.2330	25
Rm(%) x_{16}	-0.0008	0.0187	25

While testing the model's accuracy, this paper compares the predicted results of the test sample with the actual value. The equation about the percentage of accuracy is as follows:

Accuracy%=1-|predicted value – actual value|/actual value (21)

The results are in table 14. From table 14, this paper can conclude that in predicting the close stock price, the fitting effect of principal component regression is good in general. However, the lowest accuracy is 77.35%, while the highest one is 84.89%, indicating that the model is not very stable. This may result from the size of the sample in this model. Moreover, the accuracy is not perfect, which indicates that there are other significant factors with explanatory power not included in the model. Therefore, future work can conclude more samples and more influential factors.

Table 14 Evaluation Results

test close price	1	2	3	4	5
Predicted Value	18.12	10.38	2.41	9.27	5.31
Actual Value	21.6	8.9	2.06	11.99	6.25
Accuracy %	83.89%	83.33%	83.17%	77.35%	84.89%

7 Conclusion and Contributions

7.1 Conclusions

By studying the Chinese A-shared market over 2013 – 2019, this paper can address the two questions that arose in section 1.2. To answer Q1, regardless of the sentiment is bullish or bearish, the rumor causes a significant effect on the stock price before the day the rumor goes public, indicating that the source of rumor may not come from the media. When the rumor is released, although the rumor has already caused a significant effect before, stock price still reacts significantly to the rumor. It means that investment decisions established on rumors would be beneficial to investors before as well as after

they are published. As for Q2, this paper finds that after the clearance of rumor, the official clarification causes offset effect but is constrained. Clarification offsets the positive abnormal returns caused by bullish rumor significantly at a limited level. Still, after five days, it causes a positive effect like the bullish rumor did on the stock price. Moreover, clarification brings an insignificant impact on the bearish stock price. It means that the investor behaviors may not be symmetric and the market may not efficient, indicating that stock price can be manipulated.

7.2 Research Contributions

This paper makes several contributions to the literature. First and foremost, this paper uses the mean-end analysis which is one of the heuristic strategies, to construct our methodology. We start from our research questions which focus on the relationship between the investors' sentiments towards the rumor and the stock price. Because most of the existing literature identifies the investors' sentiments according to the content of the rumor which is subjective, this paper tries to classify the sentiments based on the investors' reviews, which are more direct and accurate. Therefore, this paper divides the question into several sub-goals. Firstly, the review dataset is collected by using the mean-end analysis as well. Then the sentiments in the review are identified and classified during the SMA. Thirdly, in order to do the numerical analysis, the sentiments are quantified. Finally, the quantified sentiments are introduced into the Event Study and PCA to do further analysis. The mean-end analysis is more targeted and timesaving. It is able to build a model to solve the problem in a short time using limited resources.

Secondly, this paper provides an improved methodology to study stock markets, which combines Sentiment Analysis (SMA) and Event Study altogether and the output of SMA is used as the input in the PCA. SMA allows this paper to classify investors' sentiments towards related stocks into bullish or bearish by analyzing their comments online, which express their feelings more directly, instead of the content of rumors.

Thirdly, it studies not only the effect of rumor release but also the offset effect of rumor clarification. It extends previous studies on rumor clarification. Finally, this paper provides references for market regulators who strongly advocate the disclosure of

information about unsymmetrical investor actions and the level of efficiency of the market.

7.3 Limitations and future work

There are also some limitations to this paper. The current prediction model's accuracy is between 77.35% and 84.89%. The model is not stable and there's still a gap between the predicted and actual values. To make the model more stable and narrow the gap, future work can conclude more samples and more influential factors.

Moreover, although this paper employs sentiment analysis, it is fundamental as investor sentiments are only classified into two categories. A better-designed classifier can organize the sentiments to 3 or more levels.

Based on sentiment analysis, future research can study the influence of different rumors, such as refinancing, restructuring, management scandals and so on. Research on this can have important implications for stock market regulators who strongly advocate the publication of essential firm-related information.

Acknowledgement

We are grateful to VC Research, with grant number VCR 0000017, to support this research.

References

Aggarwal, R. K., & Wu, G. (2003). Stock Market Manipulation—Theory and Evidence. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.474582

Ahern, K. R., & Sosyura, D. (2015). Rumor has it: Sensationalism in financial media. The Review of Financial Studies, 28(7), 2050-2093.

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. In W. Pedrycz & S.-M. Chen (Eds.), Sentiment Analysis and Ontology Engineering (Vol. 639, pp. 313–340). https://doi.org/10.1007/978-3-319-30319-2_13

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177-186). Physica-Verlag HD.

Black, F. (1986). Noise. The Journal of Finance, 41(3), 528–543. https://doi.org/10.1111/j.1540-6261.1986.tb04513.x

CNKI. (2019). Retrieved Nov 7th, 2019 from https://kns.cnki.net/kns/brief/default_result.aspx

Daily Trading Data. (2019). Retrieved September 7th, 2019 from https://www.gtarsc.com/SingleTable/DataBaseInfo?nodeid=4176

Davies, P. L., & Canes, M. (1978). Stock prices and the publication of second-hand information. Journal of Business, 43-56.

Diefenbach, R. E. (1972). How good is institutional brokerage research? Financial Analysts Journal, 28(1), 54-60.

Fama, E. (1991), Efficient capital markets: II, Journal of Finance, 46, 1575-1617.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. Journal of machine learning research, 9(Aug), 1871-1874.

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, *2*(3), 349-360.

Huberman, G., & Regev, T. (2001). Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. The Journal of Finance, 56(1), 387-396.

Kiymaz, H. (2001). The effects of stock market rumors on stock prices: evidence from an emerging market. Journal of Multinational Financial Management, 11(1), 105-115.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression.* New York: Springer-Verlag.

Kothari, S. P., & Warner, J. B. (2007). Econometrics of Event Studies. In B. Eckbo Espen (Ed.), Handbook of Corporate Finance: Empirical Corporate Finance (pp. 3–36). Handbooks in Finance Series, Elsevier, North-Holland.

Lafi, S. Q., & Kaneene, J. B. (1992). An explanation of the use of principal-components analysis to detect and correct for multicollinearity. Preventive Veterinary Medicine, 13(4), 261-275.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Listed company financial indicator analysis database. (2019). Retrieved September 7th, 2019 from https://www.gtarsc.com/SingleTable/DataBaseInfo?nodeid=16696

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69, 14-23.

Logue, D. E., & Tuttle, D. L. (1973). Brokerage house investment advice. Financial Review, 8(1), 38-54.

Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The journal of Finance, 25(2), 383-417.

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. The American Statistician, 36(3a), 158-160.

Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77). ACM.

Oh, C., & Sheng, O. (2011, December). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In Icis (pp. 1-19).

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71-72.

SSE. (2008). Retrieved Nov 7th, 2019, from http://www.sse.com.cn/disclosure/listedinfo/announcement/c/2008-02-26/600050_20080226_1.pdf

Su, J., Shirab, J. S., & Matwin, S. (2011). Large scale text classification using semi-supervised multinomial naive bayes. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 97-104).

Suspension and Resumption Database. (2019). Retrieved September 7th, 2019 from https://www.gtarsc.com/SingleTable/DataBaseInfo?nodeid=477

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. 16.

Yang, X., & Luo, Y. (2014). Rumor Clarification and Stock Returns: Do Bull Markets Behave Differently from Bear Markets? Emerging Markets Finance and Trade, 50(1), 197–209. https://doi.org/10.2753/REE1540-496X500111

Zhao, J. M., He, X., & Wu, F. Y. (2010). Research on Chinese stock market rumors: rumors, rumors and their impact on stock prices. Management World, (11), 48-61.